HMP Whole-Metagenome Assembly
Center for Bioinformatics and Computational Biology (CBCB)
University of Maryland College Park

Author: Mihai Pop
Version: 1.0c
Effective Date: 04/05/2011

1 Abstract

2 Introduction

This SOP describes the HMP Whole-Metagenome Annotation Pipeline run at CBCB. This pipeline generates a 'Pretty Good Assembly' - a reasonable attempt at reconstructing pieces of the organisms present in the community that are long enough to allow gene finding and other downstream analyses.

The following assumptions are made:

- This document assumes that the data being assembled are generated with Illumina, 100bp reads paired-end. While the overall assembly strategy should work with other read lengths, some of the parameters (notably the K-mer size used by SOAPdenovo) may need to be adjusted.
- The goal of the strategy outlined here is to generate a 'Pretty Good Assembly' - a reasonable attempt at reconstructing pieces of the organisms present in the community that are long enough to allow gene finding and other downstream analyses. We do not guarantee that these assemblies are the best possible given the data.
- The version of SOAPdenovo used throughout this document is 1.04 (available from http://soap.genomics.org.cn/soapdenovo.html). A newer version of this software (1.05) is now available and likely could be used instead of the version used by the HMP. We have not tested whether the newer version works with the pipeline outlined in this document. It is possible that some of the parameters we used will have to be modified, and that the outputs might be available in a format different than the one we assume in this document.

3 Requirements

3.1 Data Requirements

- Properly trimmed and pre-processed reads (see SOP on data processing for additional details)

  The reads must be organized in three files:
  - <prefix>.1.fastq
  - <prefix>.2.fastq
  - <prefix>.singleton.fastq

  Where <prefix> is a common name for the files belonging to a same experiment
  The HMP naming convention has
  <prefix> := <SRS_num>.denovo_duplicates_marked.trimmed

<SRS_num> is the Sequence Read Archive accession number for the experiment.

Note: The .1 and .2 files contain the forward and reverse reads from each mate-pair. These files must contain exactly the same number of sequences, and the corresponding sequences must occur in exactly the same order.environment, such as ("this protocol requires use of a compute cluster") should also be provided. The .singleton file contains any reads that are unmated due to the fact that their mate was excluded by one of the quality control steps performed during data processing.

- Reasonably accurate estimates of the library size used in the sequencing experiment. Note that these values cannot be simply inferred from the parameters of the DNA shearing experiment or from the gel ladder information, rather must be estimated from the actual sequences generated by the instrument. See *4.0* in the Procedure section below.

3.2 Software Requirements

SOAPdenovo version 1.04 (if using a different version see notes in *Section 2*)

4 Procedure

4.0 Estimate library sizes

Note: ideally, this step should be performed during data processing.

Prerequisites:

- compute_mates.pl Perl script
- the aligner Bowtie (bowtie-bio.sf.net)
- reference database - set of DNA sequences that are expected to be present in the sample and that are longer than the expected library size. Good options are: (i) a database of 16S rDNA sequences (if the sample being assembled is bacterial), (ii) the human genome reference (if human 'contamination' is high enough), (iii) the sequence of an organism known to be present in the sample (spiked-in control, or organism commonly found in samples of the type being analyzed). The reference database needs to be formatted for use with Bowtie using the 'bowtie-build' command.

Assumptions:

- reference database is named <DB>
- mated sequences are available in <prefix>.1.fastq and <prefix>.2.fastq

Procedure:

Author: Mihai Pop
Version: 1.0c
Effective Date: 04/05/2011

```
compute_mates.pl <DB> <prefix>.1.fastq <prefix>.2.fastq
```

Output:
The output of the 'compute_mates.pl' script has the following format (all fields TAB-delimited)

```
Number   Mean   10%Mean     Standard_deviation
50133    215    212         18.7
```

Number - number of mate-pairs that correctly match the reference
Mean - mean library size (as inferred from the alignment)
10% Mean - mean library size after excluding 10% of the outliers
Standard_deviation - standard deviation for library size estimates

Additional parameters:
The following parameters can be provided to 'compute_mates.pl'
--trim NN – exclude NN fraction of outliers
--threads NN – use NN threads
--limit NN – only align first NN reads from each file (to improve performance for large datasets)

*Note: if output files *.bout present bowtie is not run (useful if you want to play with –trim parameters)*

4.1 Prepare inputs for SOAPdenovo

Prerequisites:
- input files (.1.fastq, .2.fastq, .singleton.fastq)
- correct library size estimates (denoted with <LIBSZ> below)

SOAPdenovo requires the creation of a configuration (config.txt) file that contains, among other information, the location of the input files, and library information. The file used by the HMP starts with the following information:

```
#maximal read length
max_rd_len=75
```

Then, for each library (for each set of input files if more than one is used in the project) you need to create a block with the following structure:

```
[LIB]
avg_ins = <LIBSZ> # average insert size
reverse_seq = 0 # forward/reverse library
asm_flags = 3 # reads used for contigging and scaffolding
pair_num_cutoff = 2 # num of mates needed to scaffold across a gap
map_len = 60 # minimum length of read mapping to a contig
q1 = <prefix>.1.fastq # read1 in fastq
```

Author: Mihai Pop
Version: 1.0c
Effective Date: 04/05/2011

```
q2 = <prefix>.2.fastq # read2 in fastq
q=<prefix>.singleton.fastq
```

4.2 Run SOAPdenovo

Prerequisites

- SOAPdenovo version 1.04
- Input files (.1.fastq, .2.fastq, .singleton.fastq)
- config.txt file generated in *4.1*

Procedure

```
soapdenovo all -s <config_file> -K 25 -R -M 3 -d 1 -o <outprefix>
```

where <outprefix> is an informative name for the output files

Outputs

<outprefix>.scafSeq - The main file produced by SOAPdenovo , which contains all the scaffolds produced by the assembler in FASTA format, using Ns to separate out adjacent contigs.

Other files produced by SOAPdenovo can be discarded for the purposes of this SOP.

4.3 Extract contigs and additional scaffold information

Prerequisites

- <outprefix>.scafSeq file from SOAPdenovo
- fasta2apg.pl Perl script (note - requires Bio::Perl module)

Procedure

```
fasta2apg.pl -i <outprefix>.scafSeq -size 300 -o <outdir> -name
<outprefix>
```

Outputs

The script will generate the following files within the directory named <outdir>:

<outprefix>.scaffolds.fa - FASTA record of all scaffolds over 300 bp

<outprefix>.contigs.fa - FASTA record of all contigs contained in scaffolds over 300 bp (these contigs could be smaller than 300 bp)

<outprefix>.agp - AGP-formatted scaffold information

In addition, all scaffolds and contigs are renamed to include <outprefix> in their name, followed by a serial number.

Author: Mihai Pop
Version: 1.0c
Effective Date: 04/05/2011

4.4 Extract assembly statistics

Prerequisites

- output from fasta2apg.pl script (<outprefix>.contigs.fa, <outprefix>.scaffolds.fa)
- statistics.pl Perl script (note - requires Statistics::Descriptive and AMOS::ParseFasta modules)

Procedure

```
statistics.pl --hmp <outprefix> > <outprefix>.stats.txt
```

When using the '--hmp' option, the script expects files named <outprefix>.scaffolds.fa and <outprefix>.contigs.fa and reports both sets of statistics in the output file.

Output

The output is TAB-delimited and contains the following information:

File - name of input file
Number - total number of contigs
Total Size - total size of contigs
Min Size - minimum contig size
Max Size - maximum contig size
Average Size - average contig size
Median Size - median contig size
N50 - size of contig c, such that 50% of the total assembly size is contained in contigs larger than c.
Size @ 1Mbp - same as N50 but assuming that genome size is 2Mbp
Number @ 1Mbp - smallest number of contigs that add up to 1Mbp
Size & Number @ 2Mbp, 4Mbp, 10Mbp - same as above but for more values.

Note: the N50 is a completely useless measure in a metagenomic setting, especially if comparing multiple data-sets whose total size differs significantly.

Note: the various Size @, Number @ values are meant to estimate how contiguous are the most abundant organisms in the sample. Unlike N50s, these values can be compared across samples.

Additional parameters

If the statistics.pl script is called without the --hmp option, it assumes the input is the name of a file in FASTA format (rather than the prefix of the files generated by the fasta2apg.pl script).

The statistics.pl script automatically detects and processes files compressed with gzip or bzip2 (i.e. no need to decompress the files prior to running the script).

--nohead - do not print a header line in the output (useful if concatenating the statistics from multiple assemblies in a same spreadsheet)
--justhead - the opposite: simply print the header line
--n50base NN - set the base for computing N50 sizes to NN (i.e. assumes half the genome size is NN). This option is useful if you want to compare multiple assemblies. By setting the N50 base to the same value for all assemblies you can actually compare the N50 numbers reported by the statistics script
--limit NN - limit the computation to contigs larger than NN base-pairs.

5 Implementation

6 Discussion

7 Related Documents & References

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.01	Mihai Pop	4/5/2011	Establish SOP
1.0c		9/20/2011	Converted to standard template